



**User's Guide to the  
SAS Manitoba Multiple Tumour Data Extraction Program**

**Cancer Surveillance and Prevalence  
Analytic Network  
(C-SPAN)**

Prepared by CancerCare Manitoba for  
the Canadian Partnership Against Cancer

The **Cancer Surveillance and Epidemiology Networks** have been made possible through a financial contribution from Health Canada, provided by the Canadian Partnership Against Cancer.

The views expressed herein do not necessarily represent the views of the Canadian Partnership Against Cancer nor that of Health Canada.

---

Material appearing in this report may be reproduced or copied without permission; however, the following citation must be used:

*The Cancer Survival and Prevalence Analytic Network (C-SPAN) is an initiative of the Canadian Partnership Against Cancer, in collaboration with CancerCare Manitoba.*

## Table of Contents

<b>Accompanying Documentation.....</b>	<b>3</b>
<b>Required Variables .....</b>	<b>3</b>
<b>Description of the SAS Code.....</b>	<b>4</b>
1. <i>Create a working dataset called reg_1:.....</i>	<i>4</i>
2. <i>Define cancer site groupings based on ICDO-3 (morph3) morphologies: .....</i>	<i>4</i>
3. <i>Make additional restrictions and clean data: .....</i>	<i>5</i>
4. <i>Define exit date, death indicator and death date status: .....</i>	<i>5</i>
5. <i>Calculate exact age at diagnosis and survival time using an exact interval macro shared by Larry Ellison at Statistics Canada:.....</i>	<i>6</i>
6. <i>Define additional variables required for Paul Dickman's macro:.....</i>	<i>6</i>
7. <i>Output data to a permanent dataset: .....</i>	<i>7</i>
<b>Table 1. C-SPAN Cancer Site Definitions following the site groupings agreed upon by Cancer Surveillance and Epidemiology Networks, December 2010 .....</b>	<b>8</b>
<b>References.....</b>	<b>9</b>

## Overview

This document describes SAS code (see C-SPAN\_MB Extract Multiple Tumour\_92-06\_2011-01-31.sas) used by the Cancer Surveillance and Prevalence Analytic Network (C-SPAN) to extract and prepare data from the Manitoba Cancer Registry for use in the C-SPAN period or cohort survival programs.

The SAS code selects all primary invasive malignant cancers, excluding basal and squamous cell skin cancers, and in situ bladder cancers diagnosed during the period 1992-2006. The code also groups tumours following the same site agreed upon by the Cancer Surveillance and Epidemiology Networks in December of 2010 (see Table 1), does data cleaning, and makes data exclusions and creates new variables appropriate to survival analysis.

## Accompanying Documentation

All formats used in C-SPAN programs are defined in the accompanying format file: *C-SPAN Formats\_2011-01-31.sas*. The exact interval macro, used to compute exact age at diagnosis and survival time is in the accompanying macro: *C-SPAN\_Exact Interval Macro.sas*.

## Required Variables

The input dataset of cancer cases must include the following variables to run the program:

Variable Name	Description	Format	Example
birdatst	birth date status	\$1.	C
birthdt	birth date	yyymmdd10.	1925-09-20
deathdt	death date	yyymmdd10.	2001-06-27
dthdatst	death date status	\$1.	C
dxage	age at diagnosis	best12.	55
dxatest	diagnosis date status	\$1.	C
dxdt	diagnosis date	yyymmdd10.	1993-05-19
dxmethod	diagnosis method	\$10.	histology
gender	sex (male/ female)	\$1.	F
dxfc	postal code at diagnosis	\$6.	R3G2L8
morph3	ICDO – 3 <sup>rd</sup> edition	\$5.	94413
sphin	patient's scrambled personal health identification number	\$6.	237163
topog	ICDO site (topography)	\$4.	c619
tumourid	tumour ID number	best12.	100391644
vitalst	vital status (alive/ deceased)	\$1.	a

## Description of the SAS Code

### **1. Create a working dataset called reg\_1:**

- Read in the input registry data (*may10dat.sas7bdat* in this example)
- Define diagnosis year variable (*yydx*) using the date of diagnosis (*dxdt*)
- Select cases diagnosed from 1992 to 2006 (inclusive)
- Restrict to Manitoba residents at diagnosis, where postal code at diagnosis (*dxpc*) begins with 'R'
- Resulting SAS dataset = *reg\_1*
- Invoke Alberta's IARC conversion macro shared by Alberta Health Services (if working on CCR data, skip this step)

### **2. Define cancer site groupings based on ICDO-3 (*morph3*) morphologies:**

- Read in dataset defined above = *reg\_1*
- Define the *morph2* variable by selecting the first 4 digits of the ICDO-3 morphology field (*morph3*)
- Define the *beh* variable by selecting the behaviour code (5<sup>th</sup> digit) of the ICDO-3 morphology field (*morph3*)
- If the first four digits of the ICDO-3 morphology code (*morph2*) is less than 9590 then define cancer site groupings (*cancer\_grp*) according to topography code (*topog*) and group them using the *\$topog2f* format
- If the first four digits of the ICDO-3 morphology code (*morph2*) is greater than or equal to 9590 then define cancer site groupings (*cancer\_grp*) according to morphology (*morph2*) and group them using the *\$morph3f* format
- Select only invasive cancers and in situ bladder cancers
- Identify melanomas, basal and squamous, special cases for Non-Hodgkin Lymphomas, Leukemia, and other non-specified cancers
- Exclude mesothelioma and special brain cancer cases from organ-specific site groupings

- Exclude adolescent bone cancer
- Exclude basal and squamous skin cancer
- Resulting SAS dataset = *reg\_2*

### **3. Make additional restrictions and clean data:**

- Read in dataset created above = *reg\_2*
- Exclude cases with invalid genital organs, ie. Ovarian cancer in a patient with sex = male.
- Select cases with age at diagnosis between 15 and 99 years of age ( $15 \leq dxage \leq 99$ ).
- Exclude cases where the method of diagnosis was by death certificate only (DCO) (*dxmethod* = 'death cert')
- Exclude cases with missing sex (*sex* = ' )
- Exclude cases with missing date of birth (*birdatst* = 'Y').
- Exclude cases with vital status coded as deceased and death date equal to missing (*vitalst* = 'd' and *deathdt* = ).
- Resulting SAS dataset = *reg\_3*

### **4. Define exit date, death indicator and death date status:**

- Read in dataset created above = *reg\_3*
  - If the Vital Status field indicates a person is deceased (*vitalst* = 'd')
    - and the death date occurs before or on the end of study date (*deathdt*  $\leq$  '31Dec2006'd), then set the exit date to the date of death (*exit* = *deathdt*) and set the death status indicator to deceased (*d* = 1)
    - and the death date occurs after the end of study date (*deathdt* > '31Dec2006'd), then set the exit date to the end of study date (*exit* = '31Dec2006'd) and set the death status indicator to alive

$(d = 0)$  and set the death date status as complete ( $dthdatst = 'C'$ ) in order to use exact interval macro

- If the Vital Status field indicates a person is alive ( $vitalst = 'a'$ ), set the exit date to the date of death ( $exit = '31Dec2006'd'$ ), set the death status indicator to alive ( $d = 0$ ) and set the death date status as complete ( $dthdatst = 'C'$ ) in order to use exact interval macro
- Resulting SAS dataset = *reg\_4*

## 5. Calculate exact age at diagnosis and survival time using an exact interval macro shared by Larry Ellison at Statistics Canada:

- Read in dataset created above = *reg\_4*
- Parse date variables into components variables for use in exact interval algorithm, which returns an imputed mean age or survival time in the case of missing month or day information
- Call the exact interval macro to compute age at diagnosis
- Define the variable (*agedx*) by rounding the resulting of age at diagnosis (*dur*) divided by 365.25 to 3 decimal places
- If the integer value of *agedx* is not equal to the Cancer Registry age at diagnosis (*dxage*) then set *agedx* to the Cancer Registry age
- Exclude cases with diagnosis date prior to birth date
- Call the exact interval macro to compute survival time
- Exclude cases with death date prior to diagnosis date
- Define the new variable (*dur1*) by rounding the resulting survival time (*dur*) to the nearest integer
- Exclude cases where survival time is zero (*dur1 = 0*) and diagnosis method is autopsy (*dxmethod = 'autopsy'*)
- Resulting SAS dataset = *reg\_5*

## 6. Define additional variables required for Paul Dickman's macro:

- Read in dataset created above = *reg\_5*

- Define variable survival time (*surv\_day*) as *dur1* (*surv\_day* = *dur1*)
- Age at diagnosis is referred to as age and gender is referred to as sex in Paul Dickman's macro
- Resulting SAS dataset = *reg\_6*

**7. Output data to a permanent dataset:**

- Read in dataset created above = *reg\_6*
- Resulting SAS dataset = *reg\_multumour*

**Table 1. C-SPAN Cancer Site Definitions following the site groupings agreed upon by Cancer Surveillance and Epidemiology Networks, December 2010**

Index tumours, a selected using the above extraction process, are categorized by cancer site in the following way:

Cancer Site*	ICDO-3 Site*/Histology Type**
Oral	C000 – C148
Esophagus	C150 – C159
Stomach	C160 – C169
Colorectal (Excludes Anus)	C180 – C189, C199, C209, C260
Liver	C220
Pancreas	C250 – C259
Larynx	C320 – C329
Lung	C340 – C349
Melanoma	C440 – C449 (Histology Types: 8720 – 8790)
Breast	C500 – C509
Cervix	C530 – C539
Body of Uterus (Excludes 'uterus not otherwise specified')	C540 – C549
Ovary	C569
Prostate	C619
Testis	C620 – C629
Bladder	C670 – C679
Kidney	C649
Brain (Excludes other nervous system)	C710 – C719
Thyroid	C739
Hodgkin Lymphoma	Histology types: 9650 – 9667
Non-Hodgkin Lymphoma	Histology types: 9590 – 9596, 9670 – 9671, 9673, 9675, 9678 – 9680, 9684, 9687, 9689 – 9691, 9695, 9698 – 9702, 9705, 9708 – 9709, 9714 – 9719, 9727 – 9729 9823, 9827 - all sites except C420, C421and C424
Multiple Myeloma	Histology type: 9731 – 9732, 9734
Leukemia	Histology type: 9733, 9742, 9800 – 9801, 9805, 9820, 9826, 9831 – 9837, 9840, 9860 – 9861, 9863, 9866 – 9867, 9870 – 9876, 9891, 9895 – 9897, 9910, 9920, 9930 – 9931, 9940, 9945 – 9946, 9948, 9963 – 9964 9823, 9827 - for sites C420, C421 and C424
All Cancers	All invasive sites and in situ bladder

# Site is defined by first making the appropriate histology exclusions or inclusions.

\* ICDO-3 refers to the Third Edition of the International Classification of Diseases for Oncology, (2000)<sup>2</sup>

\*\* Histology types 9050-9055 (mesothelioma), 9140 (Kaposi Sarcoma) and 9590-9989 (leukemia, lymphoma and multiple myeloma) are excluded from other specific organ sites.

## References

1. Canadian Cancer Society's Steering Committee: *Canadian Cancer Statistics 2009*, Toronto: Canadian Cancer Society, 2009.
2. Fritz A, Jack A, Parkin DM, et al (eds.) *International Classification of Diseases for Oncology. Third Edition*. Geneva World Health Organization, 2000.